Univerzita Karlova

Filozofická Fakulta

Ústav anglického jazyka a didaktiky

Bakalářská práce

Zuzana Škutová

Rozmanitost lexika v mluveném projevu mluvčích angličtiny jako cizího jazyka jako faktor určování jazykové úrovně

Lexical variety in oral production of L2 learners of English as a factor in determining language proficiency

Praha, 2020

PhDr. Tomáš Gráf, Ph.D.

Poděkování

Ráda bych poděkovala vedoucímu své bakalářské, PhDr. Tomáši Gráfovi PhD., za jeho ochotu a trpělivost.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 07.08.2020

Abstrakt

Lexikální rozmanitost je termín popisující rozsah užité slovní zásoby v textu. Jedná se o podkategorii jazykové komplexity, která je jedním ze tří základních komponentů teorie označované CAF, která zahrnuje komplexnost (complexity), přesnost (accuracy) a plynulost (fluency). Tyto termíny jsou v lingvistickém výzkumu používány k popisu jazykové úrovně rodilých (L1) mluvčích a jsou také součástí např. rámce CEFR, jenž klasifikuje nerodilé (L2) mluvčí do kategorií na základě souhrnných jazykových dovedností. Většina zkoumání v oblasti lexikální rozmanitosti se i dnes soustřeďuje na jazyk psaný a analýza mluveného projevu bývá opomíjena, jelikož je třeba promluvy transkribovat a získaná data vytřídit. Východiskem může být použití již zkompilovaného korpusu a tato možnost byla také využita k získání dat pro tuto práci. Jedná se o subkorpus LINDSEI_CZ, jenž zaznamenává transkripce českých L2 mluvčích angličtiny, kteří byli zařazeni do jazykových úrovní B2 až C2 dle rámce CEFR. Cílem této práce bylo zjistit, zda lze v projevu mluvčích na úrovni B2 a C1 v daném korpusu pozorovat signifikantní rozdíl v lexikální rozmanitosti. Bylo zkoumáno všech dvanáct mluvčích na úrovni B2 a k nim bylo náhodným výběrem přiřazeno dvanáct mluvčích C1. Pro porovnání jejich jazykové rozmanitosti byly použity dva indexy a to TTR (Type-Token Ratio) a MTLD (Measure of Textual Lexical Diversity), které jsou dostupné v online nástroji Text Inspector. Hypotéza této práce předjímala rozdíl mezi dvěma zmíněnými skupinami mluvčích. Ani jeden z užitých indexů ovšem nepozoroval signifikantní rozdíl v lexikální rozmanitosti B2 a C1 mluvčích. Tento výsledek je možné přisoudit malému vzorku mluvčích nebo jejich velmi málo rozdílným kompetencím. Lze také spekulovat o přesnosti vyhodnocení úrovně jednotlivých mluvčí, nebo o tom, zda je lexikální rozmanitost zásadním faktorem při vnímání pokročilosti v mluveném projevu. V psaném projevu je užitá slovní zásoba klíčovým indikátorem jazykových dovedností, ale v mluveném projevu může hlavní roli přebírat plynulost.

Klíčová slova: CAF, lexikální rozmanitost, mluvený projev, L2 mluvčí, LINDSEI_CZ, korpus, TTR, MTLD, Text Inspector, CEFR

Abstract

Lexical variety (also referred to as lexical diversity) is a term used to describe the range of lexis used in texts. It constitutes a subcategory of language complexity, which is one of the three components of the CAF theory, that operates with complexity, accuracy and fluency These terms are used in linguistic research to describe language proficiency of native (L1) speakers, but they are also for example part of the CEFR framework, which classifies non-native (L2) speakers into categories based on their overall language competence. The majority of research within the area of lexical variety still focuses mainly on written language. As a result, the analysis of spoken production stays neglected. The analysis of spoken language can be more labour intensive as the data need to be transcribed and pruned before evaluation. A possible simplification would be to work with spoken language corpora that have already been compiled, which is the solution adopted to obtain data for the purpose of this thesis. The corpus used here is the LINDSEI_CZ, this sub-corpus contains transcriptions of Czech L2 speakers of English. The speakers were sorted into proficiency levels between B2 and C1 according to the CEFR standards. The aim of this thesis was to find if there is a significant difference in lexical variety between the B2 and C1 speakers in the corpus. All twelve transcriptions of B2 speakers were analysed and they were paired with twelve randomly sampled C1 speaker transcriptions. To compare their lexical variety two indexes were used-TTR (Type-Token Ratio) and MTLD (Measure of Textual Lexical Diversity), both of which are available in the online tool Text Inspector. The hypothesis of this thesis predicted a measurable difference between the two aforementioned groups. However, neither of the used lexical variety indexes observed any significant difference in the lexical variety of the B2 and C1 speakers. These results can attributed to the relatively small sample of speakers or to their highly comparable language competencies. It is also possible to contemplate, whether the proficiency levels of the speakers were correctly evaluated, or whether lexical variety is a distinguishing factor in our perception of spoken production. Arguably, in written text, lexical variety is a key factor indicating language proficiency, but in oral production, fluency could be playing the dominant role.

Key words: CAF, lexical variety, lexical diversity, spoken language, L2 speaker, LINDSEI_CZ, corpus, TTR, MTLD, Text Inspector, CEFR

Contents

1 Introduction 7
2 Key concepts and available tests 9
2.1 Dimensions of proficiency 9
2.1.1 Accuracy 9
2.1.2 Fluency 10
2.1.3 Complexity 12
2.2 Methods of measuring vocabulary 15
2.2.1 TTR and MSTTR 15
2.2.2 Vocd 16
2.2.3 MTLD 17
2.2.4 LFP 18
2.2.5 P_Lex 19
2.3 CEFR 19
3 Data 20
3.1 The learner corpus 20
3.2 Data selection 21
3.2.1 Data pruning 22
4 Method 24
4.1 TTR application 25
4.2 MTLD application 26
5 Results 26
5.1 Segmental TTR results 26
5.2 Evaluation of TTR results 27
5.3 Evaluation of MTLD results 29
5.4 TTR vs. MTLD 30
6 Discussion 31
7 Conclusion 32
8 References 34
9 Resumé 38

List of abbreviations

CAF – complexity, accuracy, fluency

CEFR – Common European Framework of Reference

IELTS – International English Language Testing System

L2 – second language

L1 – first language

LINDSEI – Louvain International Database of Spoken English Interlanguage

LFP – lexical frequency profile

MTLD – measure of textual lexical diversity

MSTTR – mean segmented type-token ratio

TOEFL – Test of English as a Foreign Language

TTR – type-token ratio

List of tables

Table 2 – Mean and segmental TTR for each speaker 27
Table 3 – Group comparison of B2 and C1 learners based on TTR 28
Table 5 - Group comparison of B2 and C1 learners based on MTLD 29
Table 6 – Speakers ordered from best to worst based on MTLD scores 29
Table 7 – Speakers ordered from best to worst based on MTLD, showing mean segmental TTR 30

1 Introduction

Lexical variety (or diversity) and other descriptions of vocabulary development are by no means a new topic of interest in the field of linguistics. Lexis has been researched for decades now, but only in the last 20 years have academics started developing a range of more sophisticated statistical methods and tools for the analysis of lexical complexity, richness, sophistication and diversity. Measuring different features of active vocabulary is a procedure that has been tested numerously in both native speaker and learner contexts. Most researchers use either the language corpus approach or compile their own language samples to tailor them to other specific research concerns of their focus. Studies of lexis are often conducted using written language data, such as student essays or other written corpora material. Recently, researchers have been attempting to expand and work with spoken language, which is much more difficult to operate with. A variety of statistical methods have been developed for the study of lexical diversity, variety and sophistication, but implementing them in research on spoken language may pose some difficulties, proving some of the methods entirely unfit or in need of tailoring for this particular use. The connection between lexical knowledge and language proficiency has also been previously explored in several articles but more material on the topic is still needed.

This is the aim of the current thesis, which attempts to explore possible correlations between lexical diversity in spoken learner language and the speakers' proficiency, the evaluation of which in itself is by no means a straight-forward process. Researchers in the field of language acquisition, didactics and teachers are faced with this issue daily. There are standardised tests available such as the Cambridge series, or the IELTS, TOEFL and more. However, there are still many questions among researchers of how accurate or appropriate this kind of testing is in determining proficiency. Especially, since proficiency is a multifaceted issue. Even though there are popularly used schematic scales for determining proficiency and neatly categorising it into levels, like the Common European Framework of Reference (Council of Europe, 2001), henceforth CEFR, we are still confronted with the problem of representativeness. As mentioned, it is not just one feature or a single linguistic capability that is measured within the framework of proficiency, making it quite a complex concept. So instead of trying to comment on all factors affecting it, this thesis focuses on a single selected feature. The central issue to this thesis is lexical variety and its connection to measuring language proficiency.

The existence of gaps in the current research has been identified and this thesis is ambitiously trying to fill in some of that space and enrich the knowledge available on the topic. It uses a spoken language corpus of L2 learners, LINDSEI_CZ (Gráf, 2017), that has been compiled already as part of a project organised at the Centre for English Corpus Linguistics at the Université Catholique de Louvain. It analyses lexical diversity using Type-Token Ratio (TTR) and the more advanced Measure of Textual Lexical Diversity (MTLD). In the end, the aim is to observe if lexical diversity is at all a reliable factor for determining language proficiency by itself, that is if the lexical diversity evaluation corresponds with the language proficiency categorisation, or if it can only be used in combination with other measures.

The next chapter of this thesis first introduces the background of the concepts that are employed in this thesis and it relates our aim to some previous research in the field. The chapter briefly looks into available means of measuring and determining lexical diversity, this is followed by a brief summarization of the conceptualisation of CEFR. The practical part starts with the introduction and the description of the data, the data selection and the pruning process. After that the next chapter focuses on utilizing TTR and MTLD as methods in measuring lexical diversity, elaborating on their strengths and weaknesses in this particular use. The following chapter provides and compares results achieved by TTR and MTLD. This is followed by a discussion, which summarizes the findings and provides a verdict on whether lexical diversity is a reliable marker of language proficiency in L2 learners. In the conclusion, limitations and future research suggestions are mentioned.

2 Key concepts and available tests

As the aim of this thesis is to compare lexical variety in spoken L2 English at two different levels of proficiency, the chapter first defines these key concepts related to language proficiency, namely complexity, accuracy and fluency. Then lexical variety is defined and the various popular techniques and tests for measuring it are discussed. Finally, the chapter also briefly explains of what use the Common European Framework of Reference can be in regard to this particular aim.

2.1 *Dimensions of proficiency*

Language proficiency is commonly described as "the degree of skill with which a person can use a language" ("language proficiency", Longman 2010: 321). Generally speaking, the more complex thoughts and ideas the person is able to express or comprehend accurately and fluently, the more proficient they are. Proficiency can be rated and measured using standardised tests and reference schemata, e.g. the Common European Framework of Reference. Operationalizing proficiency is no minor task. Usually different factors in combination are evaluated to assess the level of proficiency.

Probably the most wide-spread and common definition of this construct is offered by Housen et al. (2012) who develop the so-called Complexity, Accuracy, Fluency model, also referred to as the CAF model. It is based on the understanding that proficiency is "multi-componential in nature" (p. 3). This definition builds on proficiency having three subcategories that can be operationalized differently with respect to what research questions are explored. To provide a basic overview, each category is introduced here separately.

2.1.1 Accuracy

Accuracy can be described as "degree of deviancy from a particular norm" (Housen et al., 2012b: 4) or "the ability to produce target-like and error-free language" (Housen et al., 2012b: 2). Measuring of accuracy, however straightforward it may seem, has its issues, such as determining the standard which then would be considered the norm (Housen et al., 2012b: 3). Hammerly (1991) describes accuracy as control over the code and knowledge of the language and its systematic characteristics. There is formal standardised and centralised grammar of English, but every other aspect of the language is subject to regional varieties. Written language tends to be more conservative, but in spoken language, which is the source of data for this thesis, varieties become more apparent and divergent. In written language, native

varieties can operate with different spelling of words and overall a whole different vocabulary. When it comes to the standard in relation to oral production, apart from specific vocabulary, the varieties are distinguished by varied pronunciations. Thus, evaluating spoken accuracy on this level is often problematic.

2.1.2 Fluency

Fluency is much more complex, and it is equally complex to define. It has many subcategories and it can be viewed from a range of different perspectives. As Housen et. al (2012b) write:

> research suggests that speech fluency is a multi-componential construct in which different sub-dimensions can be distinguished, such as speed fluency (rate and density of delivery), breakdown fluency (number, length and distribution of pauses in speech) and repair fluency (number of false starts and repetitions) (p. 5)

In simpler terms, it could be summarised as "the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation" (p. 2). This definition of fluency is operationalised more easily and therefore used often in modern research, however, it is one of many. Lennon (1990) in his early description of fluency distinguishes between two types of fluency. He introduces the concepts of narrow and broad fluency. In its broad sense fluency is often used as a cover term for oral proficiency, which is an overgeneralisation. In the narrow sense it is primarily temporal, as it relates to speech being delivered at a certain "native-like" rate or tempo. Fluency in the narrow sense can be affected by a range of variables such as stress and topic of the talk. Fluency is a key component of oral proficiency, high levels of fluency in speech can override other factors and prove determining in assessing oral proficiency. That fluency is largely a temporal phenomenon and that it is evaluated often mostly on a temporal basis is supported also by Nation (1989), who evaluates fluency based on word-per-minute calculations and the number of false starts, hesitations and repetitions per 100 words.

The Longman Dictionary of Language Teaching & Applied Linguistics (2010) defines fluency as composite of "the features which give speech the qualities of being natural and normal, including native-like use of pausing, rhythm, intonation, stress, rate of speaking, and use of interjections and interruptions" (p. 222) but also notes that in second language acquisition (SLA) this term is often used to describe general proficiency, therefore being in accordance with Lennon's (1990) distinction of narrow and broad sense of fluency. Lennon's (1990) definition also mentions "native-like" pausing, it is important to not assume

that native speakers do not pause or hesitate, but the major difference of learner pauses and hesitations is in their placement (Davies 2003, Dechert & Raupach 1980, Wood 2010).

Another definition of fluency, this time cognitively based, is suggested by Segalowitz (2010). He divides the concept into three subcategories. Cognitive fluency refers to the "speaker's ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances" (p.48). This process is a race for time, the speakers need to do so efficiently and quickly, so they do not have to resort to pausing and producing disfluencies. Cognitive fluency is something that the speaker possesses. The second type of fluency Segalowitz (2010) operates with is utterance fluency. This category is the one most aligned with the above mentioned working descriptions, as it refers to the quality of the utterance and looks at its particular characteristics such as speech rate and disfluencies, namely hesitation, pausing, false starts, repetitions and self-corrections. Therefore, utterance fluency is a set of fluency characteristics of a particular speech sample. The last fluency type that Segalowitz (2010) distinguishes is perceived fluency. This relates to both of the previous types, since he defines it as a set of interferences the listener makes about the speaker's cognitive fluency based on the particular utterance fluency. It draws on impressions based on the listener's perception of the utterance that they extend to assume the speaker's general cognitive fluency. The cognitive fluency and utterance fluency may not fully correspond, since the speaker's singular performance may be affected by a number of variables like nervousness, stage fright, the topic of the discussion, whether the speech was pre-planned or not, the setting and familiarity of speaker and the listener (Housen et al. 2012b).

Other aspects which may affect fluency include the operation with formulaic language as Wood (2010) comments. He draws connection between the way learners memorise expressions, phrases and collocations as units to increased speed and ease of retrieval from memory and the use in the subsequent production, thus relating fluency to "proceduralization of knowledge" (Wood, 2010: 37). In his study of the effects of formulaic language on fluency he defines fluency as "a function of a speaker's pauses and hesitations both in temporal terms and in terms of their appropriate links with discourse pragmatics and structure" (Wood, 2010:9).

O'Brien et al (2007) in their study offer a different operalization of the term. They assert that oral fluency consists of two parts: oral ability and oral fluidity. The first being measured as total number of words and the length of the longest turn.

The latter operationalized as:

> rate of speech (words per minute), mean length of speech runs in words containing no silent pauses or hesitations greater than 400 ms, mean length of speech runs in words containing no filled pauses (ums, ahs, etc.), and longest speech run in words containing no silent or filled pauses (p.565)

The term is operationalized similarly by Freed & Segalowitz (2004), who measure fluency based on "speech rate, mean run length containing no silent pauses or hesitations greater than 400 ms, mean run length containing no filled pauses (e.g., um, ah), and longest run containing no silent or filled pauses" offering yet another view on how fluency can be conceptualised.

To include one last perspective, the concept of fluency has also been studied by Skehan (2014), who divides fluency into two sub-categories: breakdown fluency and repair fluency. Breakdown fluency is a combination of temporal factors, mainly speech rate and pausing, whereas repair fluency is concerned with language modification such as false starts, self-corrections, reformulation and repetition. As we can see there is great variety among the definitions of fluency and there is perhaps no solid central functioning definition, which reflects the fact that fluency as a concept is both multi-faceted and complicated to delineate. The definition provided at the beginning of this section from Housen et.al (2012b) implements several notions introduced by other research and is appropriate for this thesis, no overly complex subcategories of fluency need to be used here as the thesis is not specifically concerned with fluency per se.

2.1.3 Complexity

The third feature of proficiency, complexity, is perhaps the most complex one. A basic comprehensive definition of complexity provided by Housen et al. (2012b) states that "complexity is commonly characterized as the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2" (p. 2). It has been divided into two subcategories: linguistic complexity and cognitive complexity. Linguistic complexity as defined by Housen et al. (2012) "the size, elaborateness, richness and diversity of the learner's linguistic L2 system" (p. 5). Cognitive complexity is defined from the view of the speaker rather than the system. The factors that fall into the realm of cognitive complexity are different subjective factors that depend on the learner but also for example input saliency (Housen et al., 2012: 5). "Thus complexity and accuracy would relate primarily to L2 knowledge representation, or to the level of analysis of internalized L2 knowledge" whereas fluency "is primarily related to the learners' control over their linguistic L2 knowledge system

as reflected in the speed and efficiency with which they can access and implement relevant L2 information to communicate meanings in real time" (p. 6).

A different view on complexity divides it into lexical and syntactic complexity (Lahmann et al., 2015; Bayazidi et al., 2019) or sometimes grammatical and lexical complexity (Tonkyn, 2012), which is a much more practical division for the discipline of applied linguistics. Lexical complexity is the more relevant category out of the two to the present research. There are some inconsistencies in terminology in available literature, for example Read (2000) or Daller & Jarvis (2013) refer to the same concept as lexical richness, but they clearly define it in the same terms making this only a labelling difference.

The categories of lexical complexity are lexical variety, also sometimes termed lexical diversity (Tonkyn, 2012) or variation (Read, 200), lexical sophistication and lexical density (Read, 2000; Bayazidi et al., 2019; Daller & Jarvis, 2013). Housen & Bulte (2012:28) also differentiate between systemic lexical complexity, which covers the categories of lexical density and lexical variety, and structural lexical complexity which includes lexical compositionality and sophistication which is a shared subcategory of both systemic and structural lexical complexity.

2.1.3.1 Lexical Density

Lexical density is defined as proportion of content words to function words. Higher percentage of content words shows that the information is presented using concentrated and condensed language (Read, 2000; Housen & Bulte, 2012) This category was originally established by Ure (1971) to compare written and spoken texts. Johansson (2008) has demonstrated that difference between written and spoken texts are expected to occur no matter what lexical complexity category is measured, observing different significantly higher lexical density and diversity in evaluated written narratives in comparison to spoken narratives from the same speakers.

2.1.3.2 Lexical sophistication

This measure evaluates the vocabulary based on how frequent and contextually appropriate the words used in the text are (Read, 2000). The higher the number of rare words is, the more lexically sophisticated the text is. Lexical sophistication is therefore related to both breadth and depth of the vocabulary knowledge (Housen & Bulte, 2012, Kyle & Crossley, 2015) and it is often evaluated by comparing the text with some form of frequency lists based on related corpora (Young-Sook, 2018). There are several tools developed to help

automatize this process, which are introduced below. Lexical sophistication builds its relevance to proficiency on the notion that "acquisition will occur in order of frequency, suggesting that a higher proportion of lower frequency words in a learner text is a hallmark of a more elaborated mental lexicon" (Eguchi & Kyle, 2020: 382). Eguchi & Kyle (2020) mention that measures of lexical sophistication should not be limited to frequency lists, but take into consideration, polysemy, imageability, or register restrictions of the given lexical item. This approach necessitates either the incorporation of numerous indexes into the analysis or a human rater. The limitations of the use of indexes are technical, e.g. related to particular transcription style and tagging, in order for the tests to recognise the input. A major limitation of the human rater is bias. In addition to that a human rater can only work with limited samples and at a much slower pace. Rálišová (2020) compared frequency list analysis and manual evaluation of relative difficulty of the used lexis, in order to investigate whether lexical sophistication differed between the learners at B2 and C1 proficiency levels in LINDSEI_CZ. She concluded that only human manual evaluation provided results that showed some significant distinction between the two groups of learners.

2.1.3.3 Lexical variety

The key concept this thesis is working with is lexical variety, which is often also called lexical diversity. Read (2000) writes that lexical variety is the level of diversity and range of lexis that the author or speaker presents in their texts. He also claims that: "It is reasonable to expect that more proficient writers have a larger vocabulary knowledge that allows them to avoid repetition by using synonyms, superordinates and other kinds of related words." (p. 200).

Lexical variety as defined by Daller et al. (2013) is: *"*variety of vocabulary that a speaker has at his/her disposal. If the vocabulary is very small, words will be repeated often, which is an indication of low lexical diversity" (p. 196). We are offered a different distinction when Jarvis (2013) argues that lexical diversity is not a sub-category of lexical richness, the sole fact that there are such inconsistencies in terminology within one publication, goes to show how unstable these terms are. The issues of non-standardised uses of the term pose an obstacle.

There are questions to what extent learner's lexical variety as demonstrated on one occasion can be generalised to determine the overall size of the vocabulary that the learner operates with. Lexical variety is a demonstration of a theoretical capacity, which can only be measured based on singular learner performances. These performances are taken to be

representative, because there is simply no other way through which overall vocabulary can be measured, the active vocabulary is only demonstrated through the learner's production. However, especially in cases where the learner's production is not pre-mediated, we might only use the findings to approximate the overall potential lexical knowledge. Furthermore, there are great differences between written and oral production. In oral production the learner does not have time to re-evaluate their linguistic choices, and they might focus on fluency and speed rather than accuracy or complexity in communication. Written material can therefore be expected to feature richer lexis than any spontaneous oral production.

2.2 *Methods of measuring vocabulary*

Depending on the category that is being measured, there are different methods and tools available. Traditionally, oral production is evaluated by a trained evaluator based on their listening experience, as it is still done in the case of language testing in IELTS for example. However, with the rising demand for more objective evaluation, and the need to process much more data simultaneously, we observe the focus shifting towards analytical tool development. The tools are useful especially for quantitative analyses and for supplying numerical assessment of our data. There are many tools readily available for different linguistic needs. Nowadays, the biggest issue is refining the tools sufficiently both from the perspective of linguistics and statistics, so that the provided data is reliable and correctly focused. Depending on how exactly the tools and methods of vocabulary evaluation operate, we can sort them into two simple categories: internal and external measures. Internal measures work solely with the information and data provided within the text. External measures compare the data in the text with some external information or data, e.g. frequency lists in the measurements of lexical sophistication.

2.2.1 TTR and MSTTR

TTR, type-token ratio, is one of the oldest measures of vocabulary. TTR belongs into the category of text internal measures. It is very easy to calculate. The TTR is a simple ratio of types and tokens. As such, its maximum value could reach 1 if all of the words in the text were different. This measure is mentioned in virtually any article or publication that is concerned with lexical variety or any aspects of it (Laufer & Nation, 1995; Read, 2000; McKee et al., 2000; Meara & Bell, 2001; Treffers-Daller et al., 2018). The results might appear easy to interpret as it seems clear what TTR measures. Unfortunately, basic TTR is

extremely sensitive to text length, the shorter the text the more unreliable TTR is, which is understandable from just looking at the formula (Laufer & Nation, 1995). In a short text it is unlikely that we will encounter numerous repetitions, especially in written texts, so the results of TTR might show a very positive evaluation in the case of a short message of five lines. In a study with a regulated text length we can use the TTR for comparison across the texts used even if the text length is not ideal, but we cannot relate the results of the test to general values and results of other studies of texts of different text lengths. Another issue comes from a common problem in any science and that is definition of the concepts that are pivotal for this measure, in our case the issues might start with the basic definitions of the word (Laufer & Nation, 1995, Treffers-Daller et al., 2018). The other problem is that only in very short utterances could TTR reach its maximum value of 1, as much as it is virtually impossible that the ratio could actually equal to zero, which could only happen if no text was produced at all

Because of the abovementioned issues, researchers have since tried to refine TTR or offer alternative, more sophisticated measures. MSTTR stands for mean segmental type-token ratio, which is one of the slightly more sophisticated measures based directly on TTR. It can work with smaller text units, but we still have to adjust the data into units roughly the same in size as MSTTR results for different sized texts are not comparable and the results of extremely short texts are still likely to be distorted (Malvern & Richards, 2002). However, if we apply TTR or MSTTR conscious of their limitations, they can still be of use. The problem of length dependence can be fixed by standardising the length of texts used within one study, but the issues of comparability across studies remain (McKee et al, 2000)

2.2.2 Vocd

This test was discussed and implemented by a number of researchers (McKee et al. 2000; Malvern & Richard, 2002; Lai & Schwanenflugen, 2016) in an effort to overcome the limitation of the TTR/MSTTR, which they previously worked with. It employs an index known as D-index or just D. Its theoretical characterisation is provided by McKee et al (2000):

> It has been shown that a mathematical model of the curvilinear relationship between the size of a language sample and the range of vocabulary it contains can be simplified and used as a way of quantifying vocabulary diversity. (p. 335)

An important note is that it is a test still largely based on TTR and TTR needs to be completed first in order to obtain the D – index. The process is described as follows:

> The method for obtaining D values from transcripts depends on producing a graph of the way the TTR in a given transcript falls with increasing token size within the language sample, and comparing this empirical graph with the theoretical curves obtained from the mathematical model, i.e., from the equation. (Malvern & Richards, 2002:90)

There are freely available programmed computer tools such as the *vocd* (McKee et al., 2000), that can help with obtaining D – values online, simplifying the task technically. In this case of *vocd* and how it obtains the final D value, McCarthy & Jarvis (2010) provide a neat summary

> Because D is arrived at by random sampling, the value varies each time the assessment is run. Thus, to create a higher level of consistency, the procedure above is run three times, and an average D is the final output. Final values tend to range from 10 to 100, with higher values indicating greater diversity (p. 383)

On one hand, *vocd* is helpful in partial automatization of the process, on the other hand, unfortunately, it is tied to limitations related to requiring standardised input to compute the results. These issues have been encountered and described by Lai & Schwanenflugen (2016):

> There are several limitations in obtaining D scores. First, calculating D is labor intensive. The user must obtain a language sample with enough substance to calculate D and transcribe it in accordance with the specifications of the vocd program (p. 233)

This can be especially restrictive, when working with large quantities of data, say whole sections of corpora for example. The benefit of working with D instead of relying purely on TTR/MSTTR should be its relative independence of text length (Malvern & Richards, 2002), some following research has encountered issue with this again (Lai & Schwanenflugen, 2016). However, as we can see in literature on different tools for vocabulary assessment this is a pervasive issue, that is perhaps not yet completely resolved by any test and remains one of the major limitation of data evaluation and comparability.

2.2.3 MTLD

MTLD stands for Measure of Textual Lexical Diversity. This lexical variety test has been advocated by McCarthy & Jarvis (2010), who in their study of three different lexical diversity measures found MTLD to be the least length dependent and therefore the most stable across all texts. It is "a computational textual analysis tool that produces an index of lexical diversity" (McCarthy & Jarvis, 2013: 52) and its calculations are essentially based on

TTR, but utilising it in a way that should not be length sensitive. It has been incorporated into the software of Coh-Metrix as one of the lexical diversity indices. McCarthy & Jarvis (2010) explain how MTLD processes our data:

> It is calculated as the mean length of sequential word strings in a text that maintain a given TTR value (here, .720). During the calculation process, each word of the text is evaluated sequentially for its TTR. For example, . . . of (1.00) the (1.00) people (1.00) by (1.00) the (.800) people (.667) for (.714) the (.625) people (.556) . . . and so forth. However, when the default TTR factor size value (here, .720) is reached, the factor count increases by a value of 1, and the TTR evaluations are reset. Thus, given the previous example, MTLD would execute . . . of (1.00) the (1.00) people (1.00) by (1.00) the (.800) people (.667) |||FACTORS = FACTORS + 1||| for (1.00) the (1.00) people (1.00)… (p. 384)

The calculations also include partial factors as texts are unlikely to end perfectly at a completed factor. This inclusion assures no parts of the text remain unused and helps provide a more complete evaluation. The partial factor is calculated as follows:

> For example, a TTR of .887 forms 40.4% of the range between 1.00 and the full factor of .720. If a text contains 4 full factors and a remainder that has a TTR of .887, then the final factor count is 4.00 + 0.404 = 4.404 (McCarthy & Jarvis, 2010: 384)

The final MTLD value is calculated as the total number of words divided by the factor count. This process is completed twice; once in forward processing and once using backward processing, where the text is evaluated starting from the end. The final MTLD is then the mean of these two processes (McCarthy & Jarvis, 2010). The inclusion of factors and double-processing is what makes MTLD extremely length independent and stable. The test also does not discard any data that could be valuable, which can be a major perk, especially, when the data the given research is working with is limited and any losses may dramatically affect the lexical diversity values. In an effort to make this test more accessible, it has been made available also through the textinspector.com, which is the online analysis tool that this thesis utilises in MTLD calculation.

2.2.4 LFP

This measure was originally developed by Laufer & Nation (1995) to measure a different aspect of vocabulary. Lexical Frequency Profile (LFP) mainly aims to assess data within the lexical sophistication category. It is a text external measure that requires the aid of frequency lists for reference. Frequency lists are derived based on corpora and usually sort

words into different frequency tiers. The inspected text is then evaluated based on how many words from the rarer word tiers (lower frequency words) the texts features. Laufer & Nation (1995) advise to adjust the framework of evaluation for lower proficiency learners and higher proficiency learners or native speakers. This test is still dependent on text length. Laufer & Nation (1995) state that the test is mostly reliable when applied to texts over 200 hundred words in length, which can prove to be a great limitation especially in the case of low level learners, who might have trouble producing longer texts without help, making the issue of text elicitation another complex worry for the researchers in their study design.

2.2.5 P_Lex

The ambition to develop a more reliable method of measuring lexis led Meara & Bell (2001) to the development of P_Lex. The authors aim to provide a more stable method in comparison to the LFP. "P_Lex looks at the distribution of difficult words in a text, and returns a simple index that tells us how likely the occurrence of these words is." (Meara & Bell, 2001:9). P_Lex also works with the same word lists, evaluating more than lexical variety and providing information on lexical sophistication. Therefore, P_Lex would be classed as an text external measure. P_Lex produces data comparable to LFP, but acts stable when applied to the evaluation of shorter texts (Meara & Bell, 2001). On the contrary, there are some questions about its accuracy when used on longer texts.

2.3 *CEFR*

The Common European Framework of Reference for Languages (Council of Europe, 2001) is a universal language evaluation framework used across various languages. Their materials provide framework, which can be used for self-evaluation, but it also is referred to in academic scenarios. The levels range from A1 to C1 and each level is described with a number of features sorted into categories, that the learner should be able to implement and operate with when on that particular level. These categorises of learner competence are: range, accuracy, fluency, interaction and coherence. Lexical variety is most closely related to the general category of range and specifically the category which CEFR refers to as vocabulary range. CEFR thus relates the learner language level to the broadness of their vocabulary and it also mentions that starting at the B2 level the learner "can vary formulation to avoid frequent repetition" (p. 27). This wording more explicitly includes the category of lexical diversity. Some of the wording used in the categorisation within the realm of

vocabulary can be very vague especially when it comes to the last three levels B2, C1 and C2. The difference between "a broad lexical repertoire" and "very broad lexical repertoire" (p. 27), is not particularly quantifiable or imaginable in exact terms. Already at the B2 level a learner should have " a good range of vocabulary connected to his field and most general topics" (p. 27), when it comes to assessing the vocabulary through interviews focusing on general topics and more common subjects the levels might prove to be difficult to assess. Nevertheless, CEFR does create the impression that with rising proficiency the many different facets of language production become more advanced. The question arises to what extent this is true of all of the facets on all levels of proficiency, and to what extent, for example, lexical advancedness is measurable using some of the lexical tests discussed above. This is what the present thesis aims to explore.

3 Data

The hypothesis of this thesis is that there is clear measurable difference in lexical variety between learners at directly adjacent proficiency levels. For the analysis of this phenomenon two lexical diversity tests are utilised. Some of the research questions were to which extent the tests are both reliable in distinguishing between the proficiency levels and whether they produce comparable results. To answer these questions, we decided to use pre-compiled learner corpus material.

3.1 *The learner corpus*

The source of data for this thesis is the LINDSEI_CZ corpus, which is a part of a larger international project documenting advanced learner language in non-native English speakers. The Czech sub-corpus is compiled of 50 transcribed interview recordings. Each recording is an interview with one Czech learner of English. Their language proficiency levels range from B2 to C2, as the participants were mostly students of an English language and literature study programme and therefore reasonably advanced. To provide some form of framework and to make the interviews comparable with each other, the interviews follow a scenario. The interviewees picked a topic they would speak on, then they were asked about their studies and plans, and at the end they were asked to describe a series of pictures in story-like framework. They were given some time to premeditate their speech and were urged to speak as independently as possible, so as to produce larger uninterrupted segments of speech. The topics they could choose from were all personal enough so that they could easily talk about

them and connected to some emotion making them more likely to speak passionately and develop the topic naturally, even in conditions where they might be nervous.

To assess the language proficiency, after the interview, each recording was evaluated by two independent and experienced language certificate score evaluators (Huang et al., 2018). They gave scores in these categories: range, accuracy, fluency, phonological control and coherence, based on these evaluations the holistic score was then calculated and translated into the CEFR scale. In cases, where there was a marked disagreement between the two evaluations, another independent evaluation by a third expert was conducted to resolve the ambiguity. The samples are all relatively advanced learners ranging from B2 to C2, with C1 being the most populous category. There are twelve B2 learners and thirty-six C1 learners documented in the corpus. Unfortunately, only two participants fell into the C2 category, which means that there is not enough data to examine this level reliably even on our smaller scale. This thesis therefore compares the recording of B2 and C1 speakers, to establish if their lexical diversity clearly corresponds to the overall language proficiency level, i.e. if the B2 learners display a markedly lower lexical variety than the C1 learners.

3.2 *Data selection*

The C2 learner interview transcriptions were discarded, since the thesis does not look at this proficiency level as mentioned above. One of the C1 samples, which was marked very short in comparison to others by the researchers who compiled the corpus, was also dismissed as the current research is using tests sensitive to length, so having texts of lengths that are too varied would not help to establish conclusive results. To have a balanced overall sample we matched the twelve B2 texts with twelve C1 texts. To choose the C1 texts without any bias and at random we used a Python sampling script. The script was fed the thirty-five code names of the transcribed C1 recordings and was designed to randomly sample twelve C1 transcription codes, those randomly selected twelve texts are used in our analysis. However, there were still more adjustments to be done within the texts, in order for them to be ready to undergo various testing without running into any processing issues.

The recordings are of interviews, featuring the interviewer and the interviewee, whose production the research is interested in. This meant that to focus solely on the learner language the interviewer turns had to be deleted from the transcriptions, so that they would not impact the results of the lexical variety tests. Each turn in the transcription was conveniently signalled by either <A> or <B> both at the beginning and at the end of the turn,

where <A> represents the interviewer turns, which had to be erased. Running the texts through a Python script, we extracted solely the lines of the text enclosed from both sides by <B>. This method effectively removed the interviewer turns as well as the turn markers <A> and <B> which left us with only the transcribed interviewee text. After this process, the text files still contained other transcription notes such as markers for overlapping, interruptions or non-verbal reactions, e.g. laughter, coughing, lip smacking. Another category of transcription notes were periods signalling pauses, colons signalling length and pronunciation markers. Since the current research is focusing on solely lexical variety, all these notes were erased as they are deemed unimportant and possibly disturbing to our analysis. Apart from these adjustments, the text was pruned.

3.2.1 Data pruning

Included in the revision process was pruning of the transcriptions. Pruning is a regular step in research utilising transcriptions and other language data during which the data is "tidied up" before the testing itself (Lennon, 1990, Tonkyn, 2012). In the area of linguistic research, which deals with lexical variety, richness etc. in oral production, pruning usually includes the deletion of disfluencies. Disfluencies are an established blanket term, that includes repetitions, false-starts and self-corrections (Lennon, 1990) all of which were found in the transcriptions. These disfluencies are normal features (Fox Tree, 1995: 709) and compensation strategies of spoken discourse, and they are not tied directly to lexical diversity, Fox Tree (1995) defined them as "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" (p. 709). Traditionally, they are more explicitly tied to accuracy or fluency, hence the name disfluencies. That is why they were deemed acceptable to eliminate them for the purpose of the current task. Especially, repetitions where a word is repeated twice or more times in a row as a compensation strategy could detrimentally skew the results. Repetitions could grossly impact the tests results because more basic tests like the TTR, are already known to be not perfectly accurate when it comes to short texts and repetitions could dramatically lower the calculated lexical variety. If we also take into account that speakers in their oral production already use less varied vocabulary, one can see why repetitions are a big issue for our testing.

In the case of repetitions, it is important to note that not always are they markers of a negative disfluency as O'Connell & Kowal (2005) note they can have other function and can be "rhetorically or emotionally repeated word" (p. 573), in our texts this type was also present

mostly in cases of repeating “very” to intensify its effect, or in repeating “yes” or “no” several times in a row in responses. This occurrence is emphatic and functional, but its presence would definitely affect the results. Emphatically repeating “yes” three times in a row is not a same kind of indication in terms of lexical variety as using only the verbs “to be” and “to have” for the lack of more specific vocabulary, however in the results these instances would be indistinguishable. That is the reason that all immediate repetitions were eliminated.

In other cases, it could also be hard to differentiate between the types of repetitions (especially in transcription), deciding which instance is emphatic, fully intentional and lexically functional, and which is a filler to buy the speaker some time. In the end, all self-corrections, false starts, hesitations and filled pauses were manually pruned to eliminate their possible impact on the results and to prevent them from confusing the running of the tests as their form is often lexically non-standard or unfinished. Table 1 below lists the length in tokens both pruned and unpruned, detailing how many words had to be eliminated alongside with the assessed range and proficiency of the speakers. On average about 7% of the text was pruned and eliminated, with one extreme in CZ044, where 21% percent of the text needed to be pruned away. This sample was evaluated as C1 and C1- in range, suggesting that perhaps immediate repetitions and false starts do not affect our perception as negatively as could be expected. After the completion of the pruning procedure, finally we had plain text ready to be run through the automatised tests.

code	proficiency	assesed range evaluator 1	assessed range evaluator 2	length in tokens unpruned	length in tokens pruned	words pruned
CZ002	C1	C1+	C1	1977	1851	126
CZ004	B2	B2	C1	2090	1845	245
CZ007	B2	B2+	C1	1170	1098	72
CZ009	C1	C1-	B2+	2115	1981	134
CZ013	B2	C1-	C1	1740	1616	124
CZ014	B2	B2+	C1-	1644	1517	127
CZ015	B2	B2+	C1-	1807	1622	185
CZ016	B2	B2+	C1-	2239	2070	169
CZ017	B2	B2+	B2+	1910	1661	249
CZ018	B2	C1-	C1	1561	1492	69
CZ024	B2	C1-	C1-	1329	1227	102
CZ025	B2	B2+	C1-	2057	1845	212
CZ028	B2	C1-	C1	1642	1489	153
CZ029	C1	C1	C1	1870	1726	144
CZ032	C1	C1-	C1-	2028	1939	89
CZ035	C1	C1	C1+	1774	1689	85
CZ038	C1	C1-	C1	1311	1213	98
CZ039	C1	C1-	C1	1954	1777	177
CZ041	B2	C1-	C1	1290	1221	69
CZ043	C1	C1-	C1	1344	1291	53
CZ044	C1	C1	C1-	1969	1552	417
CZ045	C1	C1	C1	1471	1419	52
CZ047	C1	C1-	C1	962	910	52
CZ048	C1	C1+	C1	2282	2121	161

Table 1 – Proficiency, range, length in tokens pruned/unpruned, pruned words of chosen texts

4 Method

Our prepared pruned sample is examined using two indexes of lexical variety evaluation – TTR and MTLD, both described in the theoretical part of this thesis (see p. 15 and p.18). The first test, through which the transcription data is going to be inspected is the TTR. The TTR is a simple test, and there are several helpful tools that can calculate it widely available online and for free to the public, including the Text Inspector[1], which is used in our case. It is a useful test because it is very easy to apply, the researcher inserts the texts in the online tool and TTR is calculated for them. Furthermore, the results can be compared to a lot of other study results, that are available in journal articles and studies on similar issues. That is not to say that most TTR results are mutually comparable without limitations, but basic TTR calculations have been a feature of many research articles and so there is a plethora of data, where some parts of it are bound to be comparable to others. TTR is the first test used to establish some basis for our lexical diversity readings. TTR is used on pruned texts.

The lexical diversity results obtained from TTR are then compared with the results indicated by the more advanced evaluation provided by the MTLD, the basis of which has

[1] https://textinspector.com/workflow

already been described in the theoretical part of this thesis. Essentially, MTLD should be more text length independent and it features several statistical precautionary measures that should help it provide more stable and comparable results (McCarthy & Jarvis, 2010). The results of the two lexical diversity tests are compared to one another to show if they are in agreement or if there are discrepancies in their evaluations of the texts. The results are analysed in connection to the language proficiency levels of the learners as evaluated by the learner language experts. The overall aim is to show whether both or at least one of these lexical variety measures can clearly correlate lexical variety and language proficiency, and, specifically, if there is a significant difference shown by the lexical diversity indexes between the two proficiency levels at hand: B2 and C1. CEFR lists lexical range as contributing factor to determining language proficiency, so some analytical difference is expected to occur.

4.1 *TTR application*

TTR as one of the most basic and trivial testing options in the field of lexical diversity has been shown to be highly sensitive to the length of the text it processes, since as the text get longer words are more likely to repeat. Our corpus texts are rather long and therefore could be problematic for TTR evaluation. In an attempt, to neutralise TTR's length bias, the texts are divided into segments of approximately the same length between 200 and 250 words, where the final lexical diversity is calculated as a mean of the partial results. This modification is not a novelty as it has already been mentioned by Johnson (1944) who calls this Mean Segmental TTR. He divides the samples into segments of the exact same length, which would always result in some data loss. To preserve all data and include as much input as possible in the analysis, we instead opted to operate with a margin that would allow us to not have to discard any data leftovers. This adjustment allows for a more reliable comparison of results within this analysis as even within our reasonably balanced sample some transcriptions consist of five 200-250word long segments, whereas others had to be separated into as many as ten. The issues of comparing the results accurately to other studies are again tied to the length of the texts used there or the precautions that the researchers adopted to prevent TTR's length bias. If the researchers were using samples of approximately the same lengths or were using similar means to average out the results, then we can compare the values, otherwise there are too many circumstances that could be making the comparison inadequately unbalanced.

The thesis is aiming to determine whether there is a clear difference between the lexical variety of the B2 and C1 English learners in our sample. To examine this issue, the results across the two language proficiency levels are compared and assessed to ascertain whether a significant distinction in lexical diversity is presented in their oral production. The results are shown as total mean TTR values of the two groups and subsequently whether the difference is statistically significant is calculated using the Mann-Whitney U test as an indicator.

4.2 *MTLD application*

As in the case of TTR, the data analysed through the MTLD are the pruned versions of the oral learner texts. A big advantage of MTLD is that it relies on a more advanced analytical process and as a result it should be reasonably text length independent. On the grounds of that, the texts from our corpus are not segmented as was done previously in the case of TTR. For this test, the texts are analysed as a whole unit. MLTD already incorporates two sets of analysis and is calculated as their mean, so all MTLD scores recorded in the next chapter are based on one overall analysis. MTLD is one of the lexical diversity measures that are also available at Text Inspector portfolio. Unfortunately, Text Inspector only lets unsubscribed users analyse texts of length up to 250 words, so a subscription is necessary for the comprehensive MTLD analysis like the one required here. To determine whether there is a significant difference between the two proficiency groups based on MTLD the Mann-Whitney U test is used.

5 Results

The pruned texts of twenty-four participants were analysed using first TTR and then MTLD. Their lexical variety results are presented in the next chapter starting with TTR, then MTLD and finishing with a section that compares the results of the two tests, focusing on whether they provided corresponding results, and whether the proficiency level distinction is apparent in either of the sets of results measuring lexical diversity.

5.1 *Segmental TTR results*

As explained previously, to provide more reliable results the interviewee texts for the main section of the analysis were pruned. This step may also generate results more widely comparable to research analysing written texts, as oral production is more populated with

various disfluencies in contrast to written text. To lessen TTR's unfortunate length bias, the texts were divided into same length segments and each text's overall TTR was calculated as a mean TTR of these segments. The texts are marked with their assessed language proficiency and separated into two groups based on that.

proficiency	code	average TTR	TTR 1	TTR2	TTR3	TTR4	TTR5	TTR6	TTR7	TTR8	TTR9	TTR10
B2	CZ004	0.45	0.48	0.45	0.46	0.44	0.43	0.47	0.50	0.45	0.39	
B2	CZ007	0.45	0.50	0.51	0.53	0.55	0.55					
B2	CZ013	0.47	0.53	0.45	0.49	0.48	0.47	0.43	0.45			
B2	CZ014	0.53	0.55	0.50	0.59	0.47	0.52	0.48	0.57			
B2	CZ015	0.52	0.55	0.53	0.50	0.50	0.53	0.56	0.52	0.49		
B2	CZ016	0.51	0.50	0.50	0.52	0.47	0.51	0.48	0.53	0.56	0.53	0.49
B2	CZ017	0.52	0.49	0.54	0.57	0.52	0.53	0.51	0.50	0.46		
B2	CZ018	0.55	0.56	0.51	0.54	0.54	0.59	0.54	0.55			
B2	CZ024	0.53	0.58	0.52	0.53	0.50	0.55	0.50				
B2	CZ025	0.48	0.51	0.52	0.45	0.45	0.50	0.50	0.50	0.49	0.43	
B2	CZ028	0.48	0.52	0.48	0.49	0.45	0.45	0.49	0.45			
B2	CZ041	0.51	0.53	0.50	0.50	0.53	0.53	0.47				
C1	CZ002	0.49	0.54	0.52	0.45	0.48	0.52	0.50	0.47	0.47		
C1	CZ009	0.52	0.50	0.56	0.53	0.54	0.49	0.50	0.51	0.53	0.48	
C1	CZ029	0.48	0.49	0.46	0.47	0.52	0.52	0.40	0.51	0.44		
C1	CZ032	0.46	0.53	0.45	0.48	0.45	0.42	0.43	0.47	0.45	0.48	
C1	CZ035	0.55	0.57	0.53	0.56	0.58	0.55	0.53	0.55	0.52		
C1	CZ038	0.45	0.54	0.46	0.44	0.40	0.43	0.43				
C1	CZ039	0.49	0.49	0.47	0.41	0.49	0.53	0.52	0.48	0.50		
C1	CZ043	0.51	0.53	0.52	0.56	0.49	0.44	0.50				
C1	CZ044	0.49	0.53	0.54	0.51	0.44	0.45	0.52	0.41			
C1	CZ045	0.50	0.50	0.52	0.52	0.51	0.53	0.51	0.41			
C1	CZ047	0.46	0.48	0.44	0.47	0.45						
C1	CZ048	0.50	0.54	0.54	0.56	0.49	0.51	0.50	0.46	0.49	0.46	0.44

Table 2 – Mean and segmental TTR for each speaker

Table 2 shows the distribution of the TTR values. TTR1, TTR2, TTR3 etc. are values resulting from the analysis of the 1st chunk, 2nd chunk, 3rd chunk of the text and so on. These values were then used to calculate the mean value. Indirectly the table also shows the variance in length of the texts, with some texts only consisting of five segments and some having to be separated into as many as ten.

5.2 *Evaluation of TTR results*

Looking at the table summarising the TTR values of the texts, it becomes obvious that speakers of both proficiency levels produced language with very similar lexical variety if we simply rely on the TTR as an indicator as is shown in Table 3, which lists minims, maxims, range, mean and standard deviation of both groups.

proficiency	min	max	range	mean	SD
B2	0.45	0.55	0.10	0.50	0.03
C1	0.45	0.55	0.10	0.49	0.03

Table 3 – *Group comparison of B2 and C1 learners based on TTR*

When it comes to comparing the results of the analysis to the evaluation by the language testing experts, we can notice that for example sample CZ014 was assessed as lower level in the category of range, where the sample CZ002 was assessed as higher level for range (see Table 1) . TTR is a text internal measure, which accounts for only number of types vs token, but it does not evaluate the difficulty or complexity of the used vocabulary. If we used text external measures, such as comparing the used vocabulary to word lists, we could be able to explain such disparity in the evaluation.

Another interesting outcome of the analysis was how the mean TTRs for the two groups correlate. Surprisingly, the group mean TTR of the B2 group ended up being higher at 0.504 as the C1 TTR averaged at 0.492. Mann-Whitney U test showed that there was not a significant difference (U=37, p= 0.19) between the two groups. This result was to be expected as the two sets of values are extremely similar as shown in Table 3. In the Table 4, the results have been sorted from the highest mean TTR to the lowest to demonstrate the distribution.

proficiency	code	average TTR
C1	CZ035	0.55
B2	CZ018	0.55
B2	CZ024	0.53
B2	CZ014	0.53
B2	CZ015	0.52
C1	CZ009	0.52
B2	CZ017	0.52
B2	CZ041	0.51
B2	CZ016	0.51
C1	CZ043	0.51
C1	CZ045	0.50
C1	CZ048	0.50
C1	CZ002	0.49
C1	CZ039	0.49
C1	CZ044	0.49
B2	CZ025	0.48
C1	CZ029	0.48
B2	CZ028	0.48
B2	CZ013	0.47
C1	CZ032	0.46
C1	CZ047	0.46
B2	CZ004	0.45
B2	CZ007	0.45
C1	CZ038	0.45

Table 4 – Speakers ordered from best to worst based on mean TTR

The differences in the lexical variety as measured by the TTR are not distributed in a way that would clearly correlate with the language proficiency levels, therefore based on solely TTR our two groups are not clearly distinguishable.

5.3 *Evaluation of MTLD results*

Similarly, as TTR, the MTLD does not indicate a direct connection between lexical diversity and language proficiency within our sample. In Table 5 below, we can see that the values range more within the B2 group, which has both a lower minimum and a larger maximum, but the means of both groups are almost identical.

proficiency	min	max	range	mean	SD
B2	35.17	70.29	35.12	53.31	9.55
C1	42.31	67.22	24.91	53.00	6.91

Table 5 - Group comparison of B2 and C1 learners based on MTLD

MTLD as a length independent test does not require segmentation so the texts were analysed whole. All 24 recorded results were organised into the table in Table 6.

proficiency	code	MTLD
B2	CZ018	70.29
C1	CZ035	67.22
B2	CZ007	65.42
C1	CZ043	61.89
B2	CZ017	60.71
C1	CZ045	58.34
B2	CZ024	57.88
B2	CZ014	57.69
B2	CZ016	57.45
C1	CZ009	57.17
C1	CZ048	54.43
B2	CZ041	53.88
C1	CZ002	53.75
C1	CZ044	50.76
C1	CZ039	50.66
B2	CZ015	50.07
B2	CZ025	48.88
C1	CZ029	48.60
C1	CZ047	47.90
C1	CZ032	47.77
B2	CZ028	46.14
B2	CZ013	46.00
C1	CZ038	42.31
B2	CZ004	35.17

Table 6 – Speakers ordered from best to worst based on MTLD scores

In Table 6, our texts are sorted by their MTLD results from highest to lowest. Similarly, as in the case of the TTR, the proficiency levels do not seem to clearly correspond with the lexical diversity evaluation provided by the test, and we can see that the highest MTLD score was achieved by a learner who was evaluated as B2. Given the range of results the difference between the values the two groups as evaluated by the Mann-Whitney U test is not significant (U=37, p=0.79). In our sample we cannot correlate lexical variety as measured by the MTLD to the proficiency level classification of the speakers.

5.4 *TTR vs. MTLD*

Both the TTR and MTLD results turned inconclusive in supporting the division of the speakers into the two proficiency level groups as evaluated by the language testing experts. The two measures, however, differ in the evaluation of singular texts. The table below in Table 7 is ordered by the MLTD values from highest to lowest.

proficiency	code	MTLD	TTR
B2	CZ018	70.29	0.55
C1	CZ035	67.22	0.55
B2	CZ007	65.42	0.53
C1	CZ043	61.89	0.51
B2	CZ017	60.71	0.52
C1	CZ045	58.34	0.50
B2	CZ024	57.88	0.53
B2	CZ014	57.69	0.53
B2	CZ016	57.45	0.51
C1	CZ009	57.17	0.52
C1	CZ048	54.43	0.50
B2	CZ041	53.88	0.51
C1	CZ002	53.75	0.49
C1	CZ044	50.76	0.49
C1	CZ039	50.66	0.49
B2	CZ015	50.07	0.52
B2	CZ025	48.88	0.48
C1	CZ029	48.60	0.48
C1	CZ047	47.90	0.46
C1	CZ032	47.77	0.46
B2	CZ028	46.14	0.48
B2	CZ013	46.00	0.47
C1	CZ038	42.31	0.45
B2	CZ004	35.17	0.45

Table 7 – Speakers ordered from best to worst based on MTLD, showing mean segmental TTR

When we look at the values in the TTR column, we can see that they are out of order and do not follow the lowest to highest arrangement of the MLTD column, because the MLTD and TTR results do not correspond exactly. To determine to what extent the results of the two tests were comparable we used the Pearson product-moment correlation test, which showed a

significantly high degree of similarity ($r=0.9$, $p < 0.0001$). The results are comparable but the MTLD test is more finely grained. Where the TTR evaluates some texts as having exactly the same lexical variety the MTLD shows higher sensitivity, e.g. the last two samples in Table 7 both have TTR=0.45 but their MTLDs are 42.31 and 35.17.

6 Discussion

As observed in the Table 7, neither of our chosen indexes support the initial hypothesis of this thesis, which assumed that there would be a marked difference in lexical variety between the B2 and C1 speakers in our sample. There are multiple plausible explanations for this outcome. First of all, the LINDSEI_CZ unfortunately provides a limited sample of speakers, although their texts are sufficiently long, there is not enough of them to provide a clear characterisation of the two groups and these results can in no way be generalised.

The second possible reason for the inconclusive results could be tied to what was the focus of the evaluators in their interpretation of vocabulary range. TTR and MTLD are both text internal measure that do not take into consideration aspects of word difficulty and they do not account for such a thing as non-frequent words being a sign of a more advanced speaker. To analyse this aspect, we would have to research lexical sophistication and refer to wordlists or other measures. It is likely that the evaluators were more concerned listening for advanced words or idiomatic expressions, than for repetitions. Of course, there is also a question of whether the speaker's language level was even initially evaluated correctly.

Thirdly, it is not clear how the categories of the CAF framework affect our perception of spoken production and if lexical variety is not just a minor factor, which gets overridden by fluency or phonological control, that affect our comprehension of speech directly. It has been shown that lexical diversity indexes are reliable in differentiating between different age groups in second language acquisition (Johansson, 2008, Wu et al., 2018). Measures of lexical diversity have also been shown to present significant differences when applied to the study of task-based writing. Li (2000) reported significant differences between narrative and persuasive writing using TTR to analyse the texts, and additionally observed beneficial effects of interaction on overall language complexity, including lexical diversity. Sadeghi & Dilmaghani (2013) found that genre and topic significantly impacted lexical diversity of the texts they analysed. The effect of the topic on lexical diversity is further supported by Yu (2010), who documented differences in lexical diversity based on whether a topic was personal or impersonal. The results of these studies imply, that lexical diversity could be tied

more closely to the learner's age, genre of the text, topic or the task, rather than the learner proficiency.

Treffers-Daller et al. (2016) were aiming to correlate lexical diversity to the CEFR levels of the writers of their texts. Even though lexical diversity is perceived as a dominant distinguishing feature of written texts (Laufer & Nation, 1995, Treffers-Daller et al., 2016), the outcome of Treffers-Daller's et al. (2016) study, which utilised multiple analysis tools, was similar to the results of this thesis. Only one lemma-based method showed a relevant degree of correlation between the extracted values and the proficiency levels of the writers. Treffers-Daller et al. (2016) conclude that: " that measures of LD are very useful tools in automated analyses of students' vocabulary in essays, but it is also clear that on their own they cannot distinguish between the levels of the CEFR" (p. 322). The CEFR division into categories is also slightly problematic, as was mentioned in the theoretical part of this thesis, the respective levels do not seem to be delineated in exact terms. This problematic is documented by Rálišová (2020), who aimed to investigate the connection between lexical complexity and proficiency levels, using the same corpus as was used here, and focusing on B2 and C1 speakers. None of the quantitative methods she employed detected any difference between the two groups, and the only approach providing a significant difference was evaluation by a human rater.

7 Conclusion

It remains unclear to what extent lexical variety in oral production contributes to our perception of the speaker's proficiency or whether there is a clear correlation between the proficiency levels of speakers and the range of vocabulary they employ. Our analysis proved inconclusive using both methods and the hypothesis was not confirmed. For more accurate results that could be generalised a bigger sample of speakers would be necessary. Since this particular area of interest is largely underresearched, there is not much opportunity for comparison to studies with similar objectives or data and more research needs to be done to provide deeper insight into the problematic. Linguists have been accused of preferring written language for decades now, e.g. Linell (1982) published a whole book detailing the problematic of written language bias, but as Gilquin & De Cock (2013b) mention : "it has now become clear that the written and spoken modes present different characteristics and follow essentially different rule" (p. 6) and both modes of production should be perceived as valid and equal in the linguistic perspective. Hopefully, with technological advances the

research in spoken language will gain more momentum as new tools are being developed constantly.

8 References

Bax, Stephen. Text Inspector. textinspector.com. Last accessed on 20.07.2020.

Bayazidi, Aso & **Ansarin, Ali-Akbar & Mohammadnia, Zhila** (2019). "The Relationship between Syntactic and Lexical Complexity in Speech Monologues of EFL Learners". *Applied Research on English Language* 8(4): 473-488. https://doaj.org/article/d75e7888350e467e9d9b93eb064b2fb3

Bulté, Bram & Housen, Alex (2012) "Defining and Operationalising L2 Complexity". In Housen, Alex & Kuiken, Folkert & Vedder,Ineke eds. *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA.* Amsterdam: Benjamins Publishing Company.

Council of Europe (2001). *Common European Framework of Reference for Languages Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Dechert, Hand W & Raupach, Manfred eds. (1980). *Temporal Variables in Speech : Studies in Honour of Frieda Goldman-Eisler.* Series: Janua Linguarum, Seria Maior 86. Berlin : De Gruyter Mouton.

Daller, Helmut & Jarvis, Scott eds. (2013). *Vocabullary Knowledge: Human Ratings and Automated Measures*. Amsterdam: John Benjamins Publishing Company.

Daller, Helmut & Turlik, John & Weir, Ian (2013). "Vocabulary Acquisition and the Learning Curve". In Daller, Helmut & Jarvis, Scott eds. *Vocabullary Knowledge: Human Ratings and Automated Measures*. Amsterdam: John Benjamins Publishing Company.

Davies, Allan (2003). *The Native Speaker: The Myth and the Reality*. Series: Bilingual Education and Bilingualism, Vol. 38. Second Edition. Clevedon: Multilingual Matters.

Eguchi, Masaki & Kyle, Kristopher (2020). "Continuing to Explore the Multidimensional Nature of Lexical Sophistication: The Case of Oral Proficiency Interviews". *The Modern Language Journal* 104(2): 381-400. doi: 10.1111/modl.12637

Freed, Barbara & Segalowitz, Norman (2004). "Context, Contact and Cognition in Oral Fluency Acquisition: Learning Spanish in At Home and Study Abroad Contexts". *Studies in Second Language Acquisition* 26(2):173-199. Cambridge: Cambridge University Press.

Gilquin, Gaëtanelle & De Cock, Sylvie eds. (2013a). *Errors and Disfluencies in Spoken Corpora.* Amsterdam: John Benjamins Publishing Company.

Gilquin, Gaëtanelle & De Cock, Sylvie (2013b). "Errors and Disfluencies in Spoken Corpora: Setting the Scene" In Gilquin, Gaëtanelle & De Cock, Sylvie eds. *Errors and Disfluencies in Spoken Corpora.* Amsterdam: John Benjamins Publishing Company.

Gráf, Tomáš (2017) LINDSEI_CZ: A Corpus of Spontaneous Spoken English of Advanced Speakers. *Institute of the Czech National Corpus FF UK.* (https://wiki.korpus.cz/doku.php/en:cnk:lindsei_cz). Last accessed on 20. 07. 2020.

Hammerly, Hector (1991). *Fluency and Accuracy : Toward Balance in Language Teaching and Learning.* Series: Multilingual Matters. Bristol: Channel View Publications Ltd.

Housen, Alex & Kuiken, Folkert & Vedder,Ineke eds. (2012a). *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA.* Amsterdam: Benjamins Publishing Company.

Housen, Alex & Kuiken, Folkert & Vedder,Ineke (2012b). "Complexity, Accuracy and Fleuncy: Definitions, Measurement, Research". In **Housen, Alex & Kuiken, Folkert & Vedder, Ineke** eds. *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA.* Amsterdam: Benjamins Publishing Company.

Huang, Lan-fen & Kubelec, Simon & Keng, Nicole & Hsu, Lung-hsun (2018). "Evaluating CEFR Rater Performance through the Analysis of Spoken Learner Corpora". *Language Testing in Asia* 8: 1-17. https://doi.org/10.1186/s40468-018-0069-0

Jarvis, Scott (2013). "Defining and Measuring Lexical Diversity" In Daller, Helmut & Jarvis, Scott eds. *Vocabullary Knowledge: Human Ratings and Automated Measures*. Amsterdam: John Benjamins Publishing Company.

Johansson, Victorian (2008). "Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective". *Working Papers* 53: 61-79. https://journals.lub.lu.se/LWPL/article/view/2273/1848

Johnson, Wendell (1944). Studies in language behavior: I. A program of research. *Psychological Monographs* 56: 1–15.

Lai, Stephanie A. & Schwanenflugel, Paula J. (2016). "Validating the Use of D for Measuring Lexical Diversity in Low-Income Kindergarten Children". *Language, Speech, and Hearing Services in Schools* 47: 225-235. doi: 10.1044/2016_LSHSS-15-0028

Lahmann, Cornelia & Steinkrauss, Rasmus & Schmid, Monika (2016). "Factors Affecting Grammatical and Lexical Complexity of Long-Term L2 Speakers' Oral Proficiency". *Language Learning* 66(2): 354-385. http://dx.doi.org/10.1111/lang.12151

Laufer, Batia & Nation, Paul (1995). "Vocabulary Size and Use: Lexical Richness in L2 Written Production". *Applied Linguistics* 16(3): 307-322. doi: 10.1093/applin/16.3.307

Lennon, Paul (1990). "Investigating Fluency in EFL: A Quantitative Approach". *Language Learning* 40(3): 387-417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Li, Yili (2000). "Linguistic characteristics of ESL writing in task-based e-mail activities". *System* 28(2): 229-245. doi: 10.1016/S0346-251X(00)00009-9

Linell, Per (1982). *The Written Language Bias in Linguistics.* Linköping: University of Linköping.

Kyle, Kristopher & Crossley, Scott A. (2015). "Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application". *TESOL Quarterly* 49(4): 757-786. http://www.jstor.org/stable/43893786

Nation, Paul (1989). "Improving Speaking Fluency". *System* 17(3): 377-384. https://doi.org/10.1016/0346-251X(89)90010-9

Malvern, David & Richards, Brian (2002). **"**Investigating Accommodation in Language Proficiency Interviews Using a New Measure of Lexical Diversity". *Language Testing* 19(1): 85-104. doi: 10.1191/0265532202lt221oa

Meara, Paul & Bell, Huw (2001). "P_Lex: A Simple and Effective Way of Describing the Lexical Characteristics of Short L2 Texts". *Prospect* 16(3): 5-19. http://www.ameprc.mq.edu.au/__data/assets/pdf_file/0013/241411/Prospect_16,3_article_1.pdf

McCarthy, Philip & Jarvis, Scott (2010). "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment". *Behavior Research Meathods* 42(2): 381 – 392. doi: 10.3758/brm.42.2.381

McCarthy, Phillip & Jarvis, Scott (2013). "From Intrinsic to Extrinsic Issues of Lexical Diversity Assessment: Validation Study" In Daller, Helmut & Jarvis, Scott eds. (2013). *Vocabullary Knowledge: Human Ratings and Automated Measures*. Amsterdam: John Benjamins Publishing Company.

McKee, Gerard & Malvern, David & Richards, Brian (2000). **"**Measuring Vocabulary Diversity Using Dedicated Software". *Language and Linguistic Computing* 15(3): 323-337. doi: 10.1093/llc/15.3.323

O'Brien, Irena & Segalowitz, Norman & Freed, Barbara & Collentine, Joe (2007). "Phonological Memory Predicts Second Language Oral Fluency Gains in Adults". *Studies in Second Language Acquisition* 29(4): 557-581. https://www.jstor.org/stable/44487184

Rálišová, Diana (2020). "Measuring lexical complexity in L2 speech with word frequency lists". Diplomová práce, vedoucí PhDr. Tomáš Gráf, Ph.D. Available at Repozitář závěrečných prací UK.

Read, John (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Richards, Jack E. & Schmidt, Richard (2010) *Longman Dictionary of Language Teaching and Applied Linguistics*. Edinburgh: Pearson.

Sadeghi, Karim & Dilmaghani, Sholeh Karvani (2013). "The Relationship between Lexical Diversity and Genre in Iranian EFL Learners' Writings". *Journal of Language Teaching and Research* 4(2): 328-334. doi: 10.4304/jltr.4.2.328-334

Segalowitz, Norman (2010). *Cognitive Bases of Second Language Fluency.* New York: Routledge.

Skehan, Peter (2014). *Processing Perspectives on Task Performance.* Series: Task-based Language Teaching, Vol. 5. Amsterdam: John Benjamins Publishing Company.

Tonkyn, Alan (2012). "Measuring and Perceiving Changes in Oral Complexity, Accuracy and Fluency." In Housen, Alex & Kuiken, Folkert & Vedder,Ineke eds. *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA.* Amsterdam: Benjamins Publishing Company.

Treffers-Daller, Jeanine & Parslow, Patrick & Williams, Shirley (2018). "Back to Basics: How Measures of Lexical Diversity Can Help Discriminate between CEFR Levels". *Applied Linguistics* 39(3): 302-327. doi: 10.1093/applin/amw009

Ure, Jean (1971). "Lexical Density and Register Differentiation". In G. E. Perren & J. L. M. Trim (eds.). *Applications of linguistics*. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969: 443-452. Cambridge: Cambridge University Press.

Wood, David (2010). *Formulaic Language and Second Language Speech Fluency: Background, Evidence and Classroom Applications*. London: Continuum.

Wu, Sahng-Yu & Huang, Rei-Jane & Tsai, I-Fang (2018). "The applicability of D, MTLD, and MATTR in Mandarin–speaking children". *Journal of Communication Disorders* 77: 71-79. doi: 10.1016/j.jcomdis.2018.10.002

Young-Sook, Ryoo (2018). "Comparing Lexical Diversity and Lexical Sophistication in Korean EFL Writing: Topic and Text Length". *Multimedia-Assisted Language Learning* 21(3): 63-87. doi: 10.15702/mall.2018.21.3.63

Yu, Guoxing (2009). "Lexical diversity in writing and speaking task performances". *Applied Linguistics* 31(2): 236-259. http://dx.doi.org/10.1093/applin/amp024

9 Resumé

1. Úvod

Lexikální rozmanitost (lexical variety) a další kategorie lexikální komplexity (lexical complexity) nejsou sice novými oblastmi lingvistického zájmu, ale za posledních dvacet let zde došlo k mnoha inovacím. Díky pokrokům v technice se postupně vyvinulo několik značně sofistikovaných statistických nástrojů pro kvantitativní analýzu. Lexikální komplexita je zkoumána jak v kontextu rodilých mluvčích, tak v kontextů studentů jazyka, často za využití jazykových korpusů. Většina těchto zkoumání se ale zaměřuje převážně na psaný jazyk, i když v posledních letech začínají přibývat i studie zabývající se dosud opomíjeným mluveným jazykem. Využití metod, které jsou navrženy primárně pro analýzu psaného jazyka však není vždy jednoduché a dostupné testy musí být často přizpůsobovány.

Souvislost mezi slovní zásobou a jazykovou pokročilostí již také byla cílem zkoumání, ale je stále potřeba více materiálu na toto téma. Cílem této práce je proto prozkoumat možnou spojitost mezi lexikální rozmanitostí a celkovou jazykovou pokročilostí a zjistit, jestli je lexikální rozmanitost sama o sobě spolehlivým ukazatelem jazykové pokročilosti. Pro naše bádání je použit již zkompilovaný korpus mluveného žákovského jazyka LINDSEI_CZ (Gráf, 2017). Práce analyzuje lexikální rozmanitost za pomocí Type-Token Ratio (TTR) a poněkud propracovanějšího Measure of Textual Lexical Diversity (MTLD).

2 Klíčové koncepty a dostupné testy

V kapitole 2 je jazyková rozmanitost zasazena do kontextu teorie CAF. Tato teorie popisuje jazykovou pokročilost jako sestávající ze tří základních kategorií, jimiž jsou komplexita (complexity), přesnost (accuracy) a plynulost (fluency). Jednotlivé kategorie a jejich součásti jsou v praxi operacionalizovány různě. Přesnost je popisována jako „stupeň deviace od určité normy“ (přelože z Housen a kol., 2012b: 4) nebo jako „schopnost produkovat standardní a bezchybný jazyk (přeloženo z Housen a kol., 2012b: 2). Měření

přesnosti je proto limitováno otázkami definice jazykového standardu. Plynulost je mnohem složitější kategorií a existuje proto mnoho funkčních definic, které jsou detailně popsány v kapitole 2.1.2. Nejběžnější definicí plynulosti je asi ta vyskytující se v *The Longman Dictionary of Language Teaching & Applied Linguistics* (2010), která definuje plynulost jako „soubor rysů, které řeč dělají přirozenou a normální, včetně užívání pauz jako rodilý mluvčí, rytmu, intonace, přízvuku, tempa řeči a používání interjekcí a přerušování“ (přeloženo z p. 222). Kategorie komplexnosti také nepodléhá sjednocené definici. V kapitole 2.1.3 jsou shrnuty různé definici, z nichž pro naše účely je vhodné rozdělení na komplexitu gramatickou (grammatical complexity) a komplexitu lexikální (lexical complexity) (Tonkyn, 2012). Lexikální komplexita zahrnuje právě námi zkoumanou podkategorii lexikální rozmanitosti (lexical variety), vedle lexikální hustoty (lexical density) a lexikální sofistikovanost (lexical sophistication). Lexikální rozmanitost popisuje rozsah slovníku mluvčího a předpokládá se, že pokročilejší mluvčí mají více rozvitou slovní zásobu (Read 2000), což je základním předpokladem naší hypotézy.

V části 2.2 je popsáno několik dostupných testů na měření vyprodukovaného lexika. Pro analýzu v této práci byly vybrány dva – TTR a MTLD. TTR je základní a jednoduchý test, ale bohužel citlivý na délku textu. V případech, kdy nejsou všechny texty jednotně dlouhé se často buď vybírají z textu úseky o stejné délce, nebo se texty musí rozčlenit do segmentů, u kterých se vypočítá jejich TTR a následně se zprůměruje. MTLD je propracovanější a dle studií by mělo být na délce textu nezávislé (McCarthy & Jarvis, 2010). V části 2.3 je popsána konceptualizace CEFR, včetně toho, jaké kategorie hodnotí a do jakých úrovní studenty jazyka řadí.

3. Data

Zdrojem jazykových dat je v této práci žákovský korpus mluveného jazyka LINDSEI_CZ (Gráf, 2017). Korpus obsahuje 50 transkribovaných nahrávek rozhovorů se žáky na úrovních

B2 až C2. Porovnávány jsou úrovně B2 a C1. K 12 nahrávkám B2 bylo náhodně vybráno 12 nahrávek C1. Jelikož se jednalo o přepisy rozhovorů, bylo třeba eliminovat výroky tazatele, kromě toho taky všechny transkripční poznámky zachycující například smích, kašel apod. Tímto vznikl text, který zachycuje pouze promluvu žáka. Dalším krokem bylo redukování (pruning) disfluencí. V mluveném jazyce se mnohem častěji vyskytují repetice (repetitions), sebeopravy (self-corrections) nebo falešné začátky (false-starts). Tyto jevy by mohly zkreslit výsledky testů, hlavně TTR, které je už tak citlivé na délku textu. Zopakování slova 3x za sebou, by mohlo velmi snížit výsledné skóre. Tyto jevy nevypovídají o slovní zásobě studenta, ale spíše o plynulosti jeho projevu, jelikož často slouží jako kompenzační strategie pro získání více času při plánování dalšího projevu, a proto mohly být pro naše účely odstraněny. Průměrně se toto odstranění týkalo asi 7 % textu, ale v jednom případě šlo až o 21 % textu, viz Tabulka 1 (Table 1) v kapitole 2.3.1.

4. Metoda

Pro testování je použit nástroj Text Inspector (textinspector.com), který umožňuje analyzovat data za pomocí různých testů, včetně TTR a MTLD. Pro testování za pomocí TTR byly jednotlivé transkripce rozděleny do segmentů v rozsahu 200-250 slov. Toto opatření bylo použito pro prevenci zkreslení výsledků, kvůli různým délkám textů v korpusu, jež by znemožnilo spolehlivé srovnání. Každý segment byl otestován zvlášť a TTR skóre nahrávky je průměr těchto segmentálních výsledků. Následně byla porovnána data skupiny B2 a C1 za pomoci Mann-Whitney U testu. Pro užití MTLD nebylo třeba texty segmentovat, protože MTLD by mělo být na délce textu nezávislé. Výsledky MTLD skupin B2 a C1 byly opět porovnány za pomocí Mann-Whitney U testu.

5. Výsledky

Výsledky segmentálního TTR jsou doloženy v tabulce 2 (Table 2) v části 5.1, kde jsou vidět jednotlivé TTR hodnoty segmentů, průměrné TTR každého mluvčího, ale také do kolika

segmentů museli být texty rozděleny. V tabulce 3 (Table 3) v části 5.2 jsou porovnány výsledky obou skupin mluvčích, jsou zde ukázána maxima, minima, rozpětí a směrodatné odchylky. U obou skupin se jedná o velmi podobné hodnoty. Mann-Whitney U test následně potvrzuje, že mezi těmito skupinami není signifikantní rozdíl ($U=37$, $p=0.79$).

Pro vyhodnocení za pomocí MTLD nebylo potřeba texty segmentovat a výsledky následně průměrovat. V tabulce 6 (Table 6) v sekci 5.3 jsou výsledky srovnány od nejlepšího k nejhoršímu na základě MTLD, a při pohledu na rozložení hodnot, se opět neukazuje jasné rozdělení do pokročilostních skupin, nejlepší hodnocení dokonce získává student na úrovní B2. V tabulce 5 (Table 5) v téže sekci, jsou opět srovnány minima, maxima, rozpětí a směrodatné odchylky skupin B2 a C1. Skupiny vykazují velmi podobné výsledky, což je potvrzeno užitím Mann-Whitney U testu, který nenachází signifikantní rozdíl ($U=37$, $p=0.79$) mezi těmito skupinami.

V části 5.4 jsou srovnány výsledky TTR a MTLD, a je zde patrné, že se vyhodnocení neshodují jednoznačně, viz tabulka 7 (Table 7). Pearsonův koeficient korelace ovšem dokládá, že výsledky vykazují vysokou úroveň podobnosti ($r=0.9$, $p < 0.0001$). MTLD se zdá být přesnější a citlivější, jelikož v případech, kdy TTR některé texty hodnotí stejně, MTLD mezi nimi diferenciuje. MTLD se tedy zdá být vhodnějším testem, jednak díky nezávislosti na délce textu, ale i přesnější rozlišovací schopnosti.

6. Diskuze

Oba testy vykazují obdobné výsledky, ale práce s MTLD, díky zabudovaným statistickým opatřením byla snazší, tyto výsledky jsou lépe srovnatelné s jinými studiemi a také rozlišují mezi drobnějšími rozdíly. Při testovaní se v obou případech nepotvrdila hypotéza, že lexikální rozmanitost je spolehlivým ukazatelem obecné jazykové pokročilosti. Existuje několik faktorů, které toto mohly zapříčinit. Použitý vzorek je celkem malý,

dohromady se jedná pouze o 24 mluvčích. Nabízí se taky otázky ohledně správnosti předešlého vyhodnocení pokročilosti mluvčích.

Je ale také možné, že lexikální rozmanitost v mluveném projevu není klíčovým faktorem v hodnocení pokročilosti mluvčích. Je popsáno, že výsledky testů lexikální rozmanitost jsou spolehlivým faktorem pro rozlišování věku mluvčích (Johansson, 2008, Wu et al., 2018). Dále byly také zachyceny spojitosti mezi žánrem nebo účelem textu a lexikální rozmanitostí (Li, 2000; Sadeghi & Dilmaghani, 2013). Yu (2010) také poukazuje na vztah mezi osobním a neosobním tématem textu a lexikální rozmanitostí.

Naopak Treffers-Daller a kol. (2016), kteří se také snažili usouvztažnit úroveň mluvčích dle CEFR škály a lexikální rozmanitost jejich projevu za pomocí kvantitativních testů, také nenašli jasné spojení. Stejně tak Rálišová (2020), která se zaměřovala na lexikální komplexitu a pracovala se stejným korpusem, nenašla spojitost mezi jazykovou úrovní a lexikální komplexitou vyhodnocenou kvantitativními testy. Rálišová (2020) dokumentuje spojitost pouze, je-li komplexita hodnocena lidským vyhodnocovatelem. Což nabízí otázku, nakolik jsou kvantitativní testy spolehlivé.

7. Závěr

Zůstává otázkou, do jaké míry přispívá lexikální rozmanitost k našemu vnímání jazykové pokročilosti mluvčích. Naše analýza se ukázala být neprůkazná a hypotéza nebyla potvrzena. Pro zobecnění výsledků by ale bylo potřeba pracovat s větším vzorkem. Mluvený jazyk je stále z velké části neprobádaný, a proto se nenaskytuje ani mnoho příležitostní pro srovnání výsledků. Už Linell (1982) zmiňuje problematiku upřednostňování psaného jazyka. S vědeckými a technologickými pokroky se snad usnadní práce s mluvenými daty a zájem o tuto oblast se v budoucnu rozroste.